A computational scientist's perspective on current and future hardware architectures

Karl Rupp[a] **and Josef Weinbub**[b]

[a] *Institute for Microelectronics, TU Wien*
[b] *Christian Doppler Laboratory for HPTCAD, Institute for Microelectronics, TU Wien*

Power constraints prohibit further increases in clock frequency and thus single-threaded performance [1]. As a remedy, hardware vendors equip their processors with multiple cores to further increase the overall computational power provided. However, raw computational power can only be leveraged if data can be accessed and moved quickly. On the other hand, collective operations such as global reductions either within a single compute node or across a compute cluster are typically limited by latency, which cannot be reduced indefinitely due to fundamental physical limits. To successfully design algorithms and implementations for current and future supercomputers it is mandatory to have a solid understanding of these limits. Most importantly, this requires computational scientists to use parallel algorithms with medium- to fine-grained parallelism already on the node-level. Finding and exposing such levels of parallelism is, however, often difficult and subject to ongoing research in many application areas [2].

In this talk we evaluate current and future hardware architectures to aid the design of the forth generation of the Vienna Scientific Cluster (VSC-4). Our focus is on limits on strong and weak scalability, synchronization and data transfer latency, arithmetic intensity, as well as available programming models for typical hardware used in high performance computing. We will present benchmark results to quantify these limits and explain the application areas for which the respective benchmarks are relevant.

Overall, our findings confirm that central processing units (CPUs) are best suited for general purpose workloads and are most attractive for investing in long-term code modernization efforts. Graphics processing units (GPUs) and Intel's many-integrated-core (MICs) devices provide a narrow –yet attractive– sweet spot for applications that are either bound by the floating point operation rate or memory bandwidth (Fig. 1). However, today's availability of software in science and engineering which can efficiently make use of such many-core platforms is limited as is the experience and skillset among the developers, warranting only a limited availability of supercomputers powered by GPUs and MICs. To date, other accelerator platforms as well as ARM-based hardware cannot be recommended for VSC-4 because of either specialized use cases or lack of maturity.

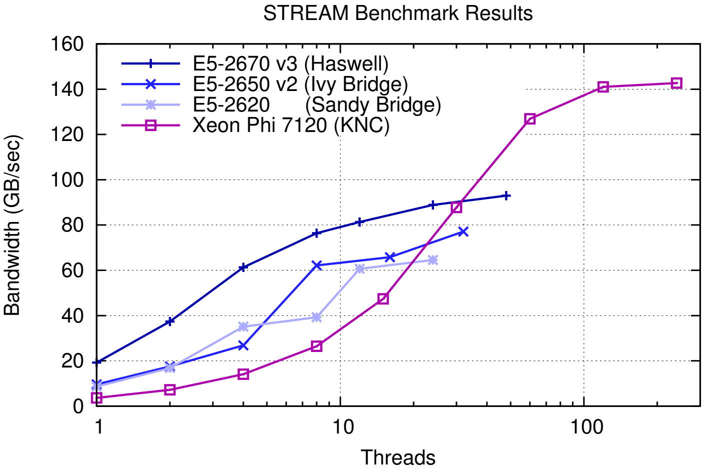

Fig. 1: STREAM benchmark results obtained for current Intel hardware. While four to eight threads are sufficient to achieve a large fraction of peak memory bandwidth on CPUs, MICs require at least 64 active threads to achieve high memory bandwidth.

References

[1] Villa, O., Johnson, D. R., O'Connor, M., Bolotin, E., Nellans, D., Luitjens, J., Sakharnykh, N., Wang, P., Micikevicius, P., Scudiero, A., Keckler, S. W., and Dally, W. J., Proc. SC'14, 830 (2014).

[2] Demidov, D., Ahnert, K., Rupp, K., and Gottschling, P., SIAM J. Sci. Comp., **35**, 453 (2013).